New CSC computing resources

Atte Sillanpää, Nino Runeberg
CSC – IT Center for Science Ltd.

Outline

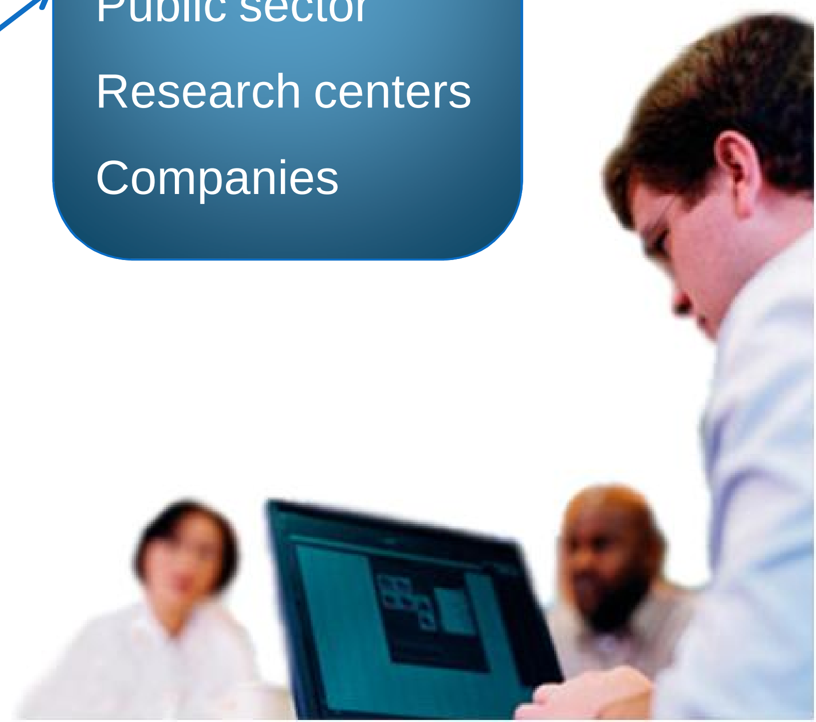
- CSC at a glance
- New Kajaani Data Centre
- Finland's new supercomputers
  - *Sisu* (Cray XC30)
  - *Taito* (HP cluster)
- CSC resources available for researchers



CSC's Services

- Funet Services
- Computing Services
- Application Services
- Data Services for Science and Culture
- Information Management Services

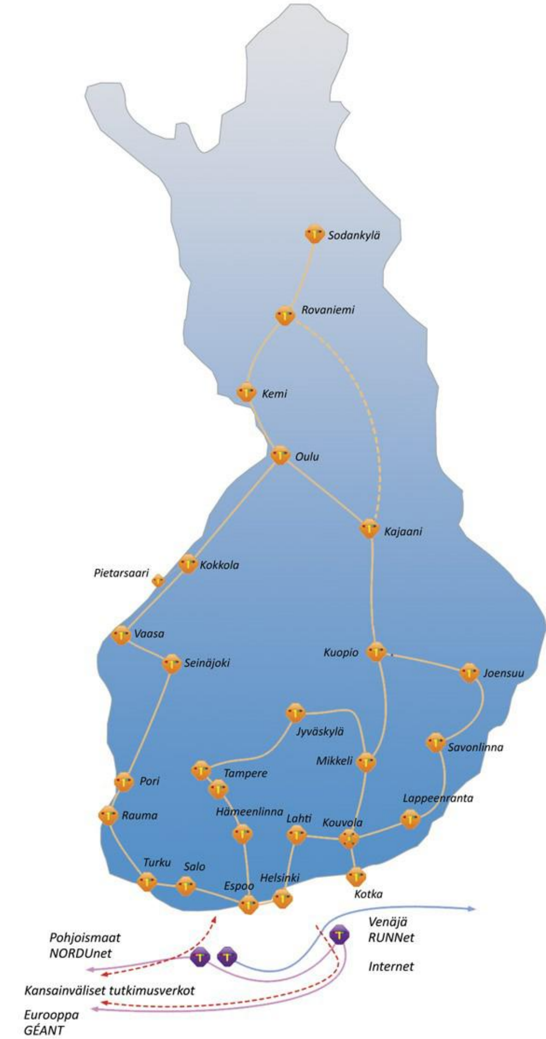
Universities
Polytechnics
Ministries
Public sector
Research centers
Companies



FUNET and Data services

- ***FUNET***
  - Connections to all higher education institutions in Finland and for 37 state research institutes and other organizations
  - Network Services and Light paths
  - Network Security – Funet CERT
  - eduroam – wireless network roaming
  - Haka-identity Management
  - Campus Support
  - The NORDUnet network

- ***Data services***
  - Digital Preservation and Data for Research
    - Data for Research (TTA), National Digital Library (KDK)
    - International collaboration via EU projects (EUDAT, APARSEN, ODE, SIM4RDM)
  - Database and information services
    - Paituli: GIS service
    - Nic.funet.fi – freely distributable files with FTP since 1990
    - CSC Stream
    - Database administration services
  - Memory organizations (Finnish university and polytechnics libraries, Finnish National Audiovisual Archive, Finnish National Archives, Finnish National Gallery)

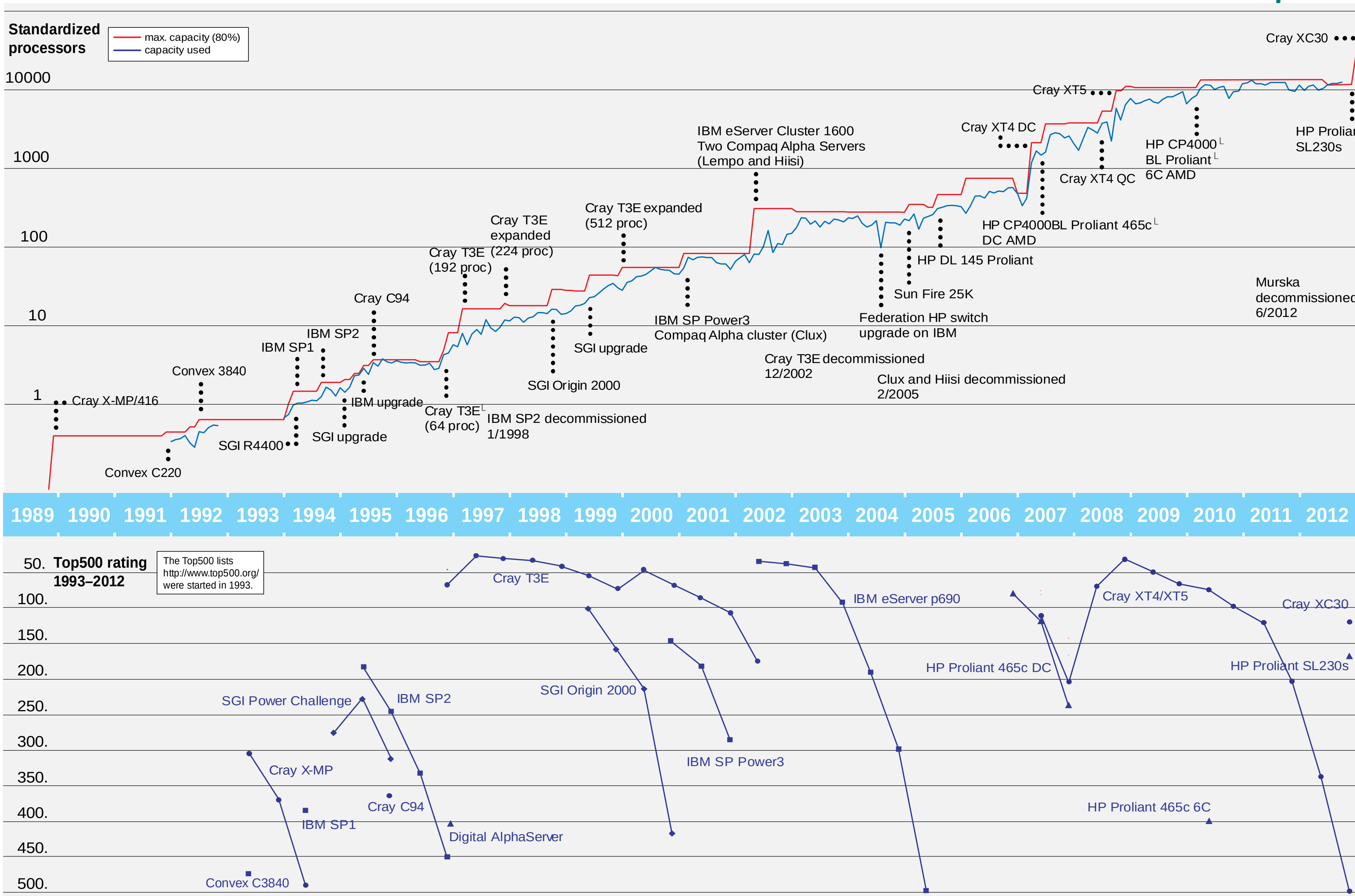


Current HPC System Environment

Name	Louhi	Vuori
Type	Cray XT4/5	HP Cluster
DOB	2007	2010
Nodes	1864	304
CPU Cores	10864	3648
Performance	~110 TFlop/s	34 TF
Total memory	~11 TB	5 TB
Interconnect	Cray SeaStar 3D Torus	QDR IB Fat tree

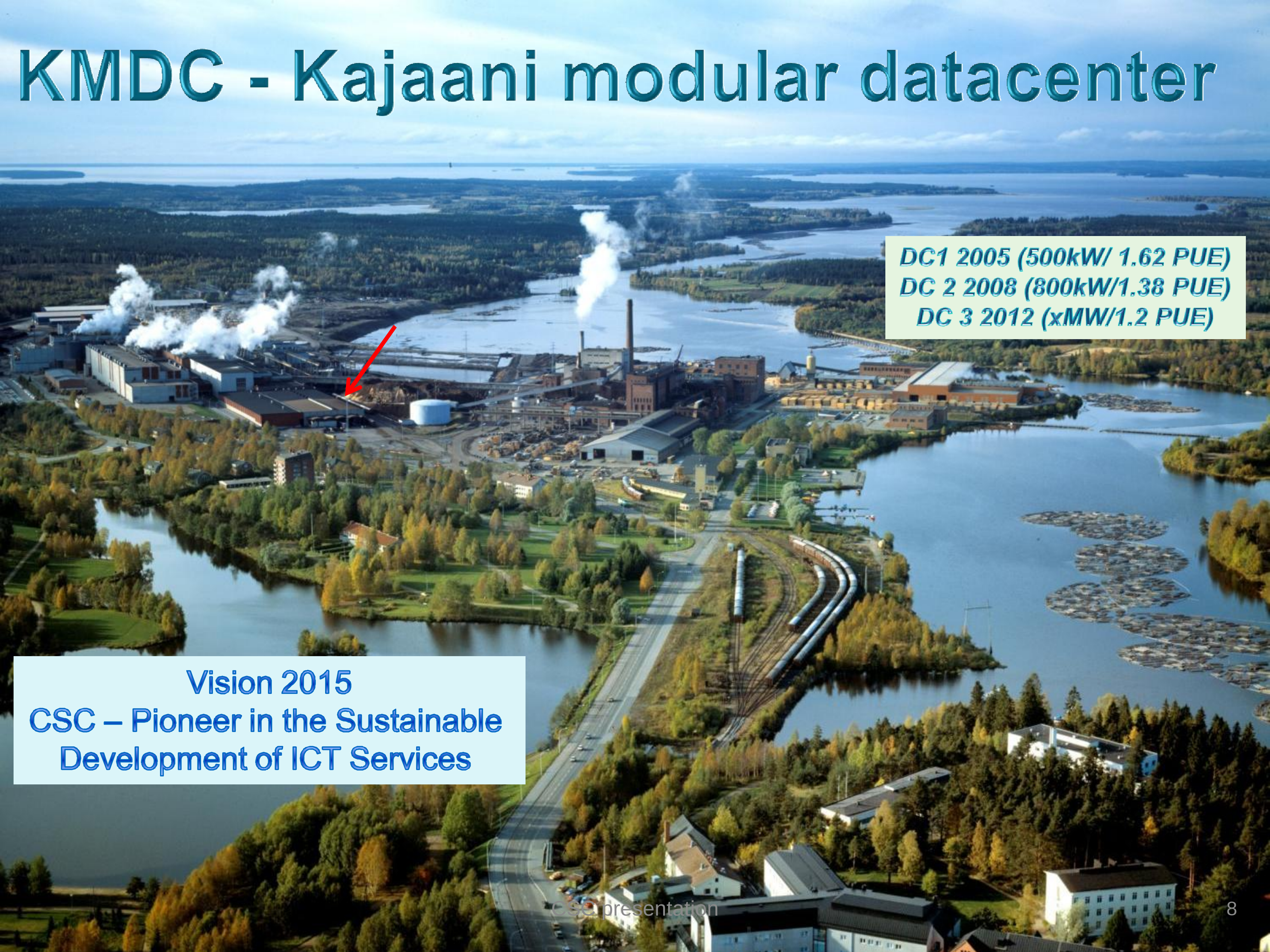


CSC Computing Capacity 1989–2012



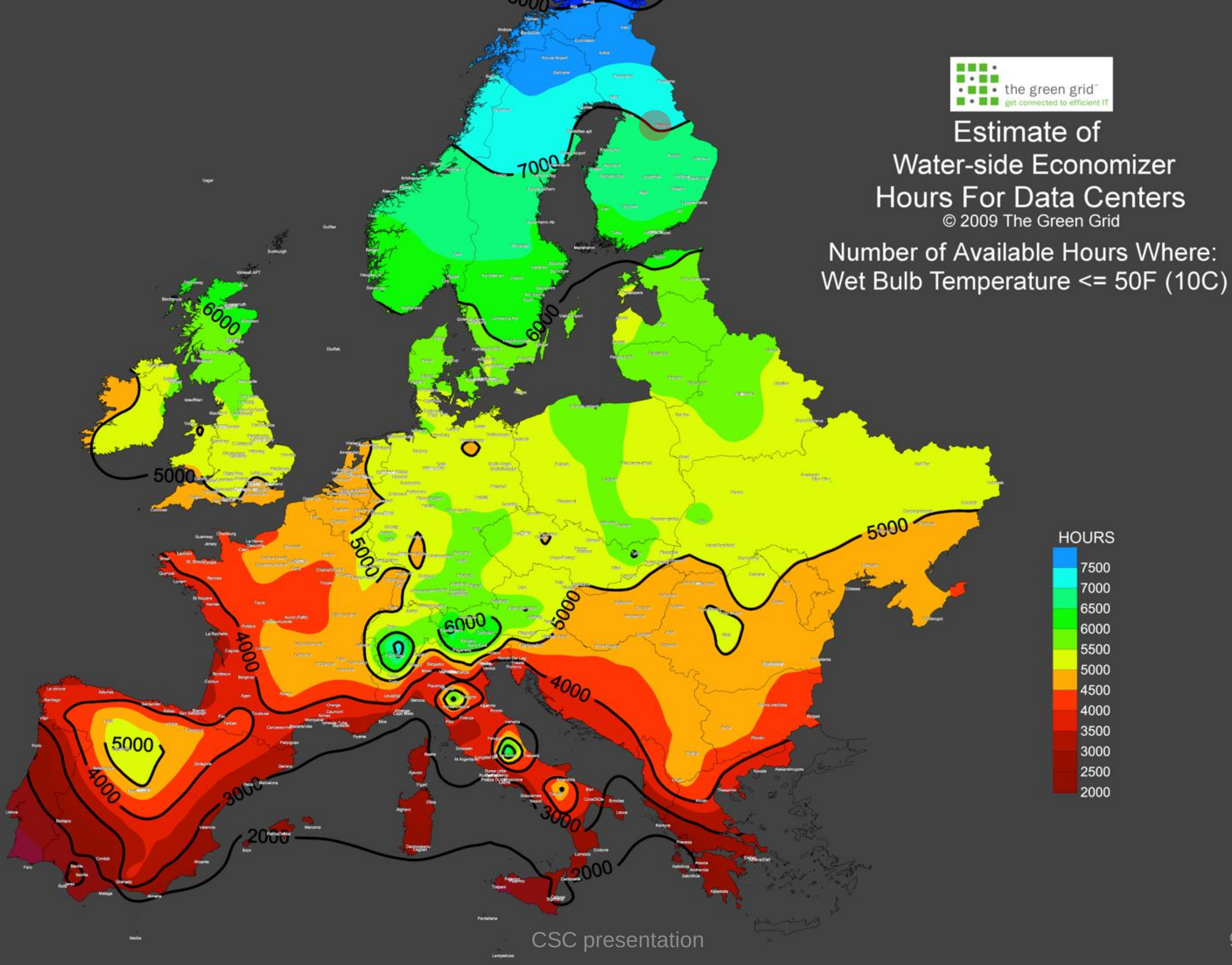
THE NEW DATACENTER

KMDC - Kajaani modular datacenter

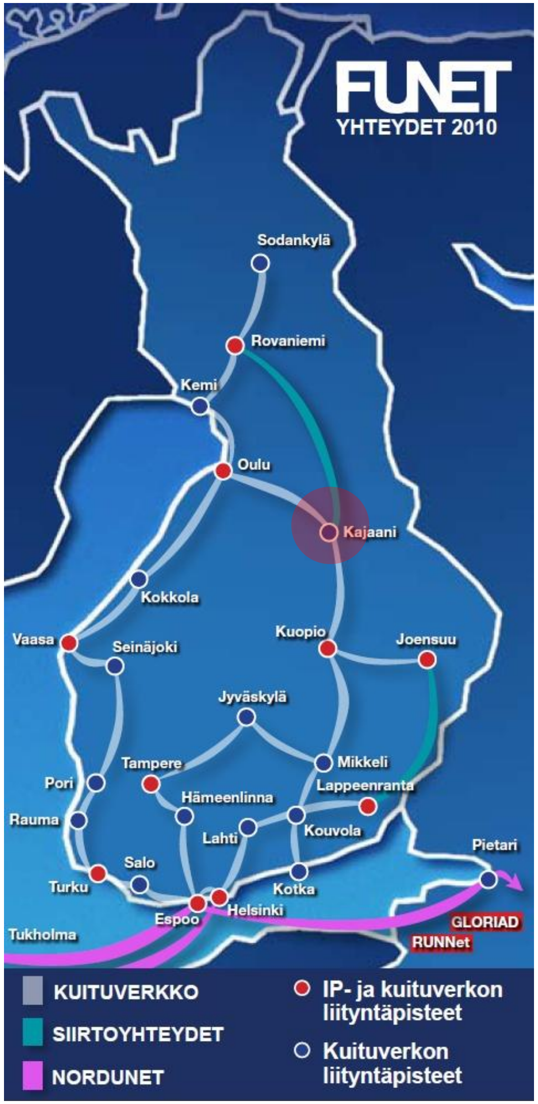


DC1 2005 (500kW/ 1.62 PUE)
DC 2 2008 (800kW/1.38 PUE)
DC 3 2012 (xMW/1.2 PUE)

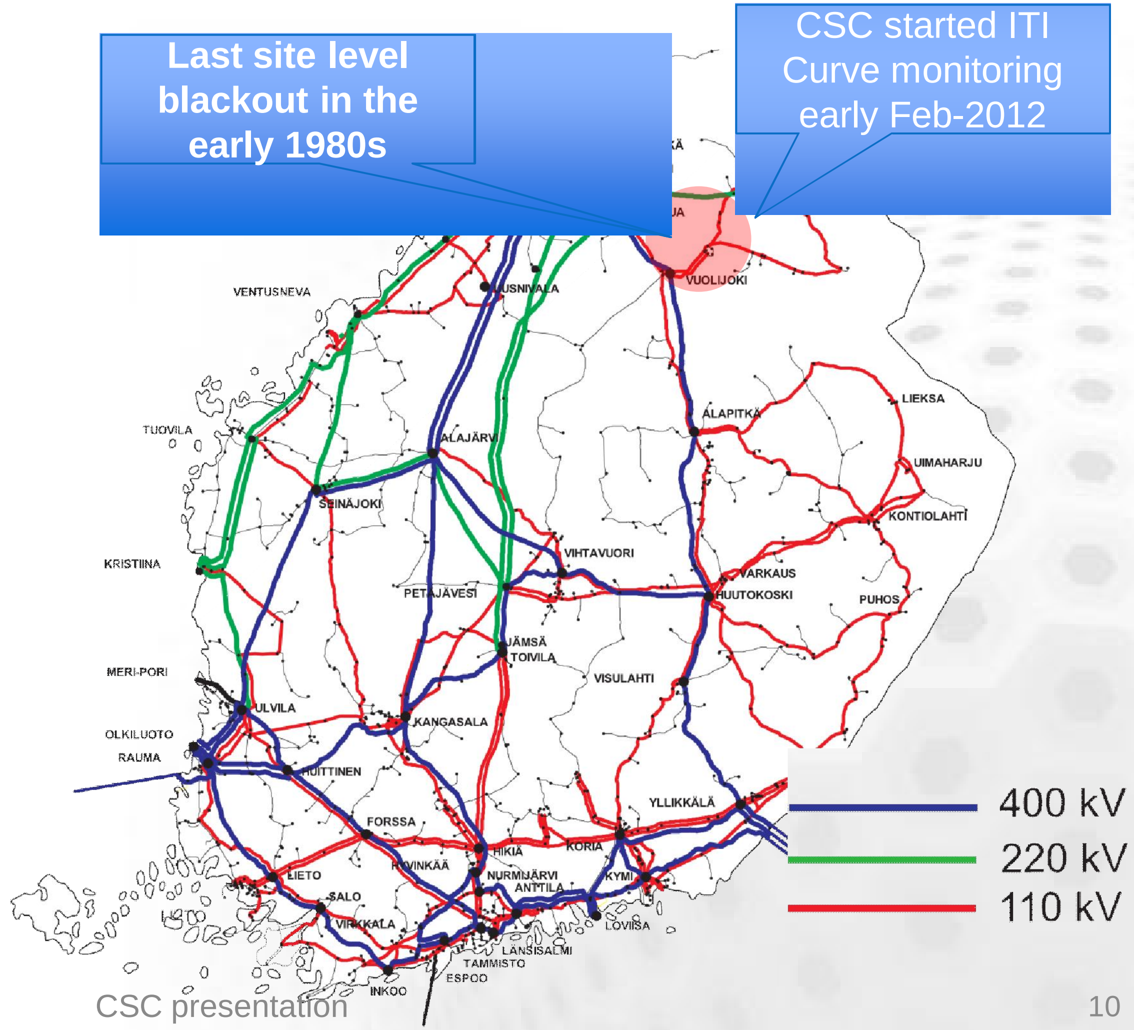
Vision 2015
CSC – Pioneer in the Sustainable Development of ICT Services



the green grid
get connected to efficient IT
Estimate of
Water-side Economizer
Hours For Data Centers
© 2009 The Green Grid
Number of Available Hours Where:
Wet Bulb Temperature <= 50F (10C)
HOURS
7500
7000
6500
6000
5500
5000
4500
4000
3500
3000
2500
2000



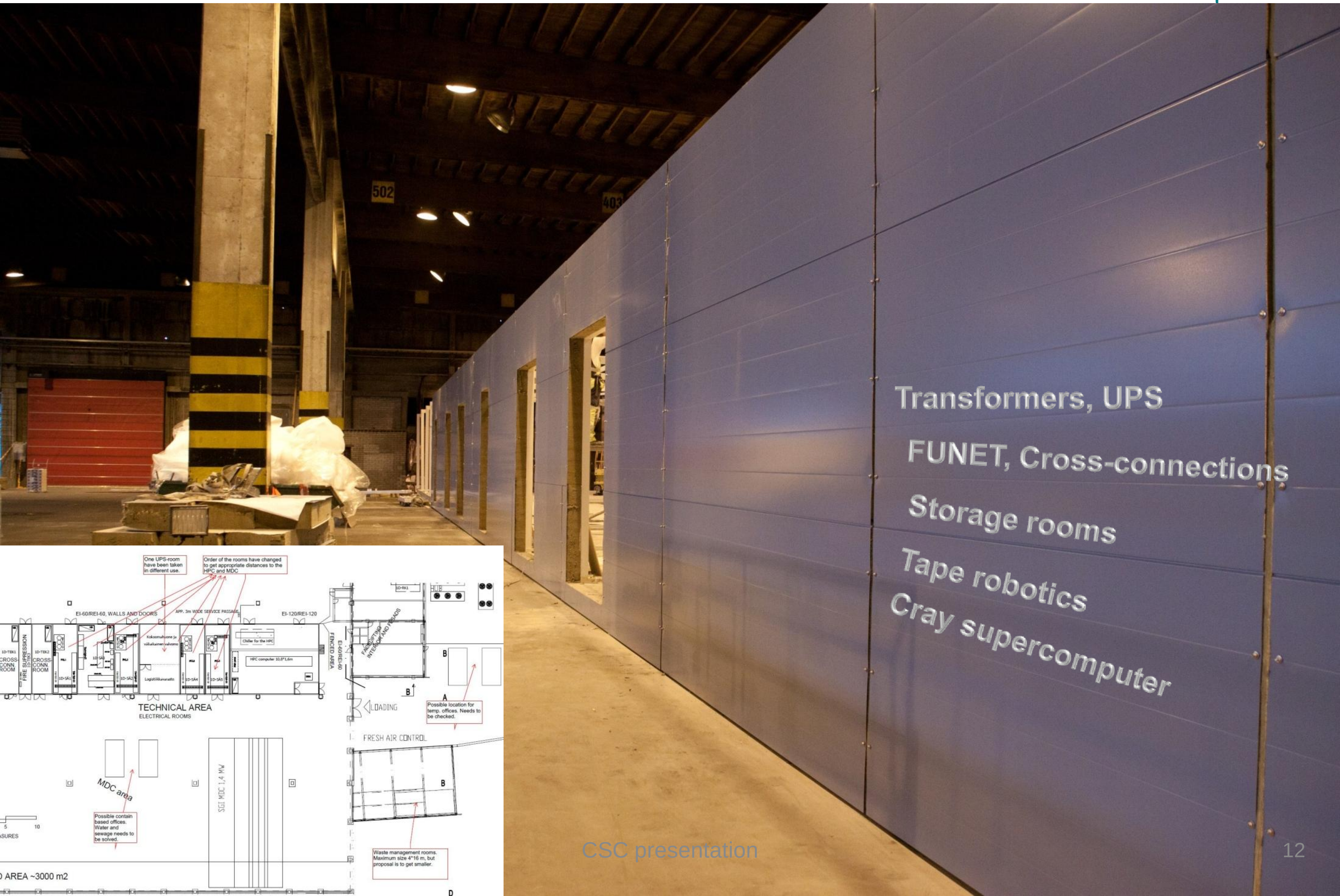
Power distribution (FinGrid)



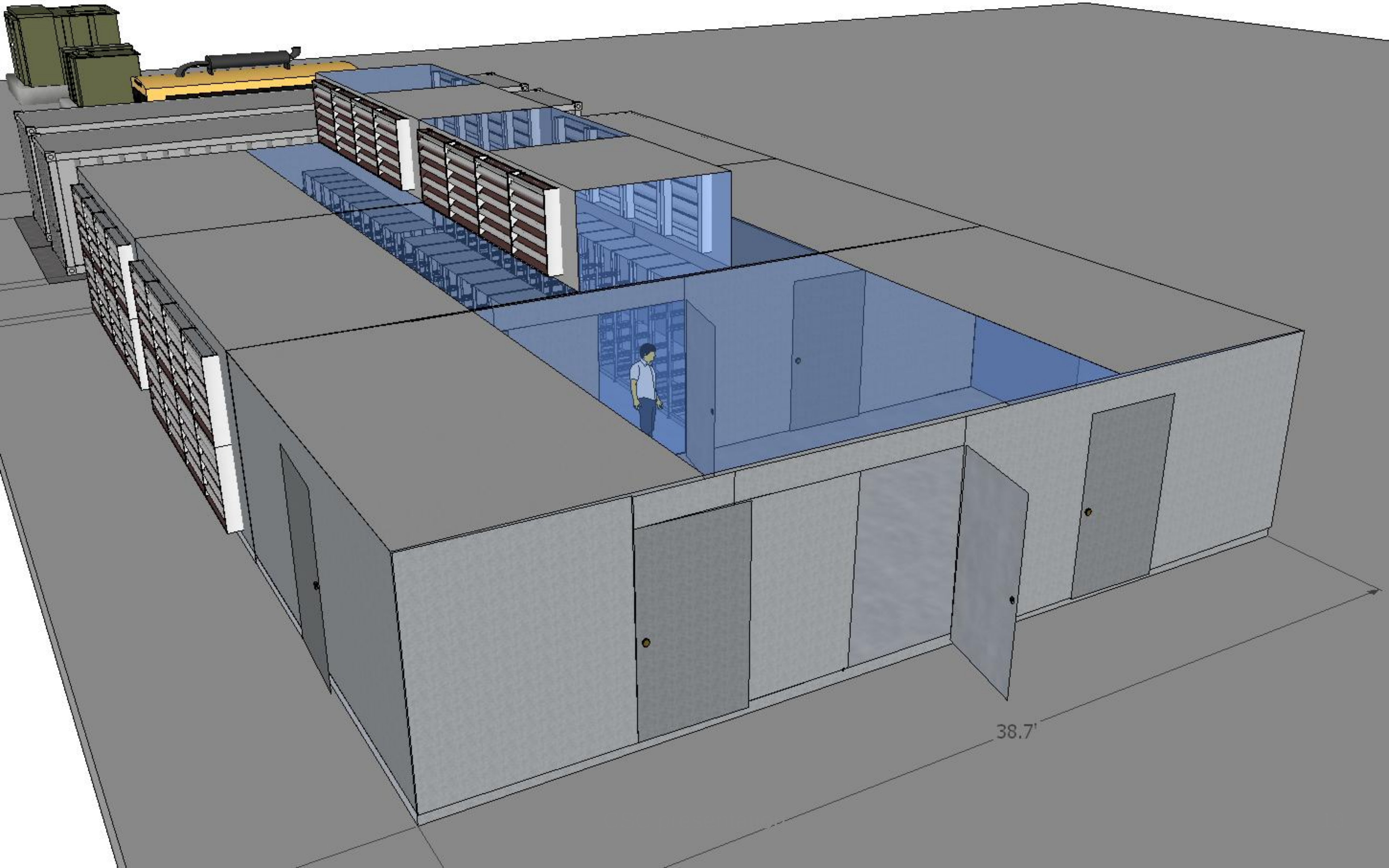
The machine hall



Sisu (Cray) supercomputer housing



SGI Ice Cube R80, hosting Taito (HP)



SGI Ice Cube R80

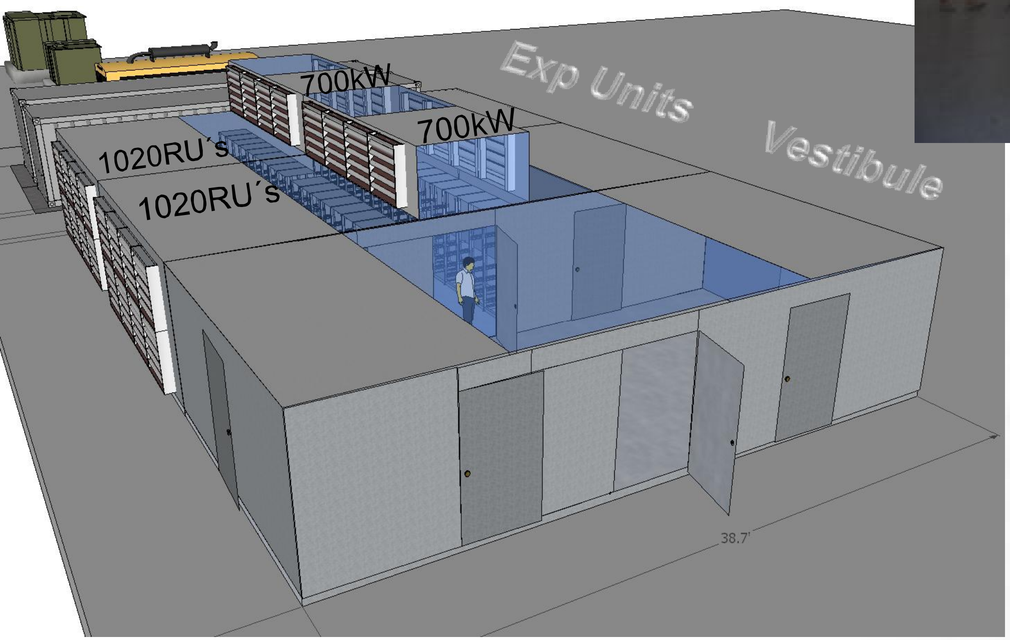


SGI MDC

Starting with one head unit (Vestibule) and two expansion modules ; extra capacity can be increased by introducing more expansion units.

Thanks to dozens of automated cooling fans, the energy needed for cooling can be adjusted very accurately as IT capacity is increased gradually.





Internal temperature setpoint 27°C (ASHRAE) and occasionally ASHRAE tolerated (27-30°C) during possible summer heat waves.

As long as outdoor temperatures are less than 28°C, Unit does nothing but free cooling. During heat waves extra water and some chillers possibly needed.

During winter, the exhaust (warm) air is re-circulated to warm up the incoming air.

Data center specification

- 2.4 MW combined hybrid capacity
- 1.4 MW modular free air cooled datacenter
  - Upgradable in 700 kW factory built modules
  - Order to acceptance in 5 months
  - 35 kW per extra tall racks – 12 kW common in industry
  - PUE forecast < 1.08 ($pPUE_{L2,YC}$)
- 1MW HPC datacenter
  - Optimised for Cray super & T-Platforms prototype
  - 90% Water cooling

CSC NEW SUPERCOMPUTERS

Overview of New Systems

	Phase 1		Phase 2	
	Cray	HP	Cray	HP
Deployment	December	Now	Mid 2014	Mid 2014
CPU	Intel Sandy Bridge 8 cores @ 2.6 GHz		Next generation processors	
Interconnect	Aries	FDR InfiniBand (56 Gbps)	Aries	EDR InfiniBand (100 Gbps)
Cores	11776	9216	~40000	~17000
Tflops	244 (2x Louhi)	180 (5x Vuori)	1700 (16x Louhi)	515 (15x Vuori)
Tflops total	424 (3.6x Louhi)		2215 (20.7x Louhi)	

IT summary

- Cray XC30 supercomputer (Sisu)
  - Fastest computer in Finland
  - Phase 1: 385 kW, 244 Tflop/s
  - Very high density, large racks
- T-Platforms prototype (Phase 2)
  - Very high density hot-water cooled rack
  - Intel processors, Intel and NVIDIA accelerators
  - Theoretical 400 TFlops performance

IT summary cont.

- HP (Taito)
  - 1152 Intel CPUs
  - 180 TFlop/s
  - 30 kW 47 U racks

DataDirect NETWORKS

- HPC storage
  - 3 PB of fast parallel storage
  - Supports Cray and HP systems

Features

- Cray XC30
  - Completely new system design
    - Departure from the XT* design (2004)
  - First Cray with Intel CPUs
  - High-density water-cooled chassis
    - ~1200 cores/chassis
  - New ”Aries” interconnect



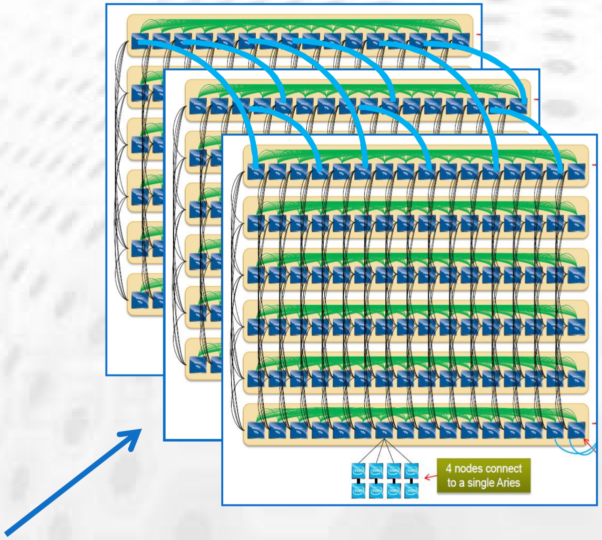
- HP Cluster
  - Modular SL-series systems
  - Mellanox FDR (56 Gbps) Interconnect





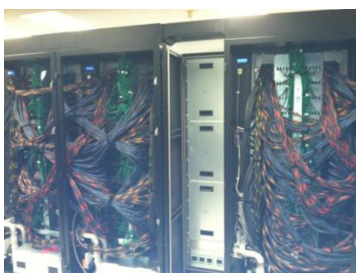
CSC new systems: What's new?

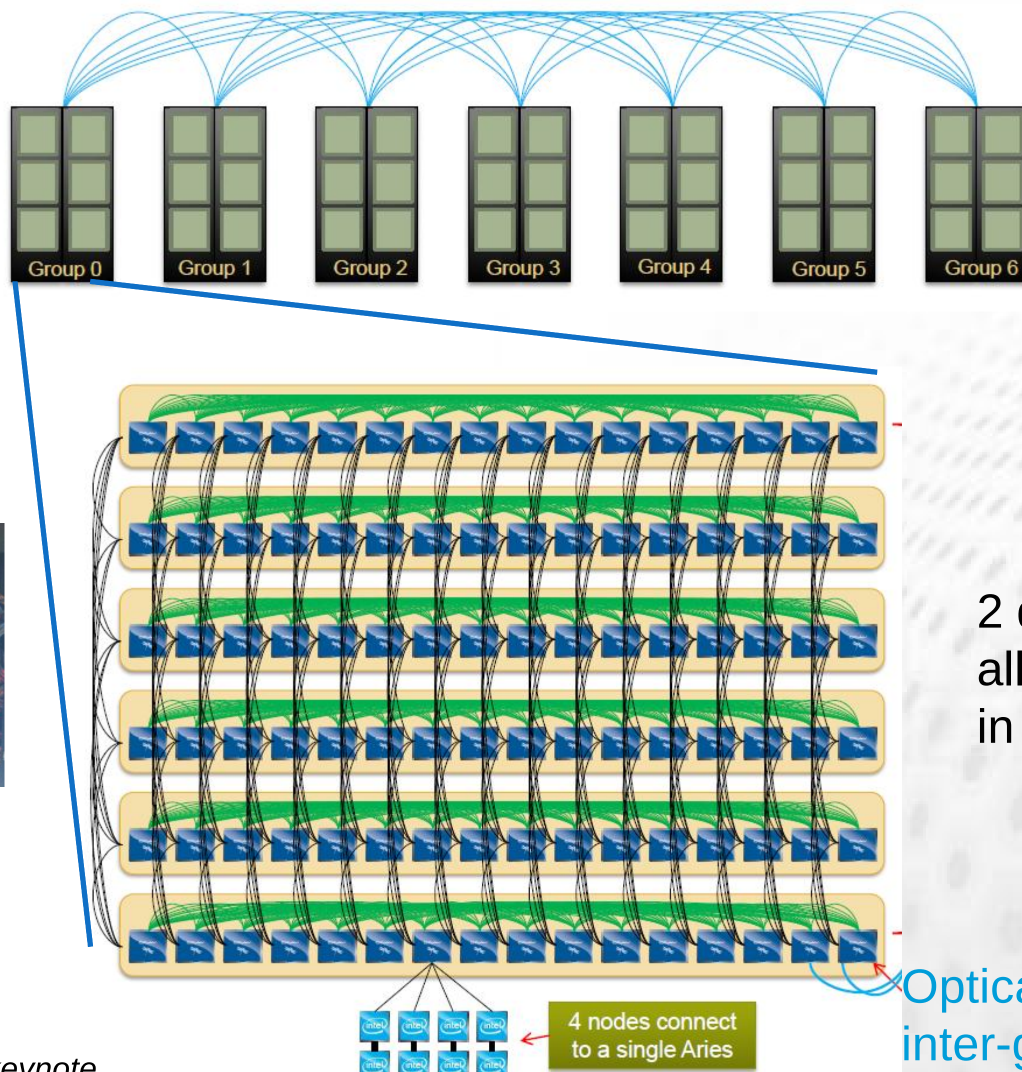
- Sandy Bridge CPUs
  - 4->8 cores/socket
  - ~2.3x Louhi flops/socket
    - 256-bit SIMD instructions (AVX)
- Interconnects
  - Performance improvements
    - Latency, bandwidth, collectives
    - One-sided communication
  - New topologies
    - Cray: "Dragonfly": Islands of 2D Meshes
    - HP: Islands of fat trees



Cray Dragonfly Topology







All-to-all network between groups

2 dimensional all-to-all network in a group

Optical uplinks to inter-group net

Source:
Robert Alverson, Cray
Hot Interconnects 2012 keynote

Cray environment

- Typical Cray environment
- Compilers: Cray, Intel and GNU
- Debuggers
  – Totalview, tokens shared between HP and Cray
- Cray mpi
- Cray tuned versions of all usual libraries
- SLURM
- Module system similar to Louhi
- Default shell now bash (previously tcsh)

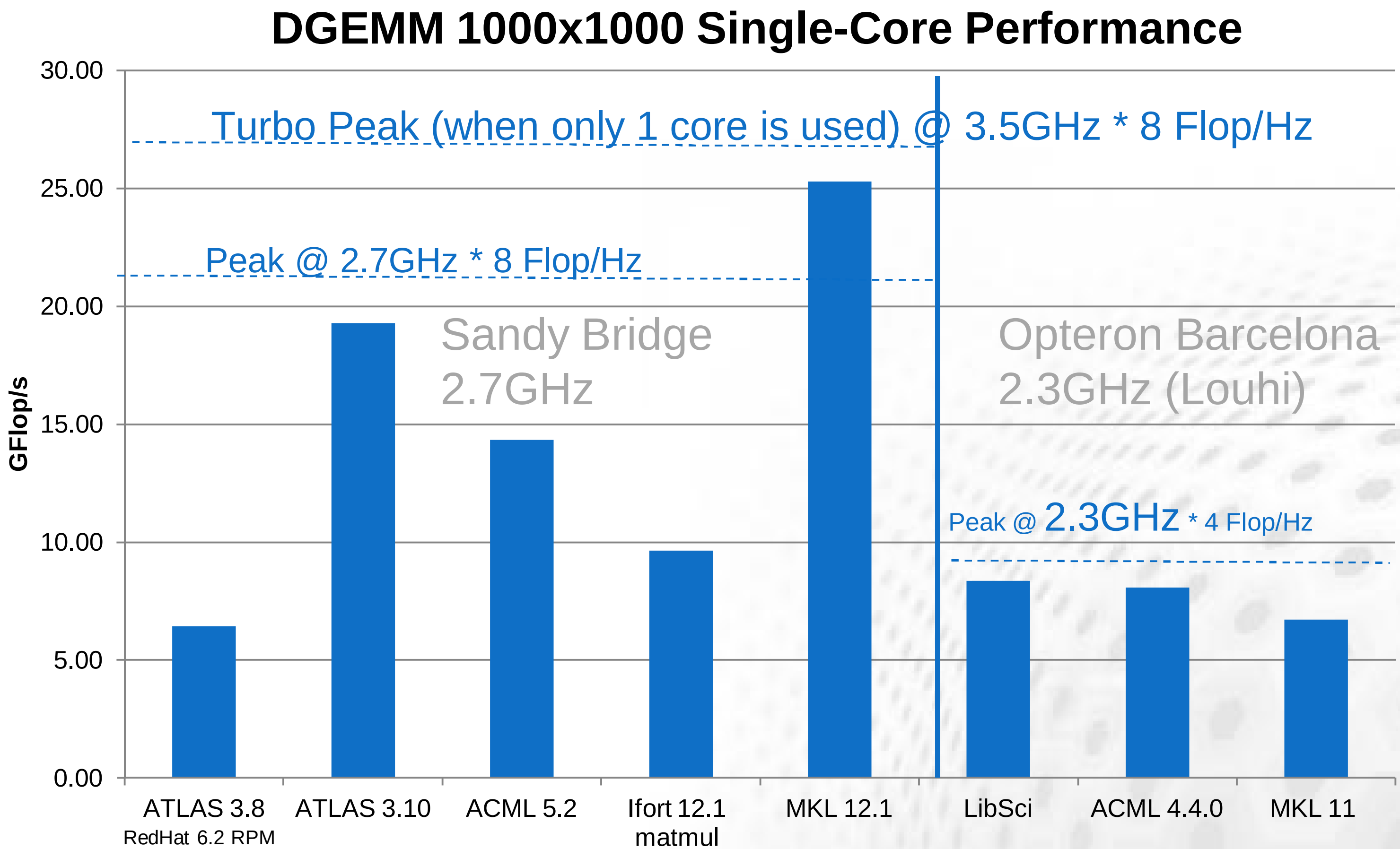
HP Environment

- Compilers: Intel, GNU
- MPI libraries: Intel, mvapich2, OpenMPI
- Batch queue: SLURM
- New more robust module system
  - Only compatible modules shown with *module avail*
  - Use *module spider* to see all
- Default shell now bash (used to be tcsh)
- Disk system changes

Core development tools

- Intel XE Development Tools
  - Compilers
    - C/C++ (icc), Fortran (ifort), Cilk+
  - Profilers and trace utilities
    - Vtune, Thread checker, MPI checker
  - MKL numerical library
  - Intel MPI library (only on HP)
- Cray Application Development Environment
- GNU Compiler Collection
- TotalView debugger

Performance of numerical libraries



MKL the best choice on Sandy Bridge, for now.

(On Cray, LibSci will likely be a good alternative)

Compilers, programming

- Intel
  - Intel Cluster Studio XE 2013
  - http://software.intel.com/en-us/intel-cluster-studio-xe
- GNU
  - GNU-compilers, e.g. GCC 4.7.2.
  - http://gcc.gnu.org/
- Intel can be used together with GNU
  - E.g. gcc or gfortran + MKL + IntelMPI
- mvapich2 MPI-library also supported
  - It can be used that Intel or GNU

Available applications

- Ready:
  - Taito: Gromacs, NAMD, Gaussian, Turbomole, Amber, CP2K, Elmer, VASP
  - Sisu: Gromacs, GPAW, Elmer, VASP
- CSC offers ~240 scientific applications
  - Porting them all is a big task
  - Most if not all (from Vuori) should be available
    - Some installations upon request
  - Do you have priorities?

Porting strategy

- At least recompile
  - Legacy binaries may run, but not optimally
  - Intel compilers preferred for performance
  - Use Intel MKL or Cray LibSci (not ACML!)
    - ***http://software.intel.com/sites/products/mkl/***
  - Use compiler flags (i.e. -xhost -O2 (includes –xAVX))
- Explore optimal thread/task placement
  - Intra-node and internode
- Refactor the code if necessary
  - OpenMP/MPI workload balance
  - Rewrite any SSE assembler or intrinsics
- HPC Advisory Council has best practices for many codes
  - http://www.hpcadvisorycouncil.com/subgroups_hpc_works.php
- During (and after) pilot usage, share your makefiles and optimization experiences in the wiki
- http://software.intel.com/sites/default/files/m/d/4/1/d/8/optaps_cls.pdf
- http://software.intel.com/sites/default/files/m/d/4/1/d/8/optaps_for.pdf

Modules

- Some software installations are conflicting with each other
  - For example different versions of programs and libraries
- Modules facilitate the installation of conflicting packages to a single system
  - User can select the desired environment and tools using module commands
  - Can also be done "on-the-fly"

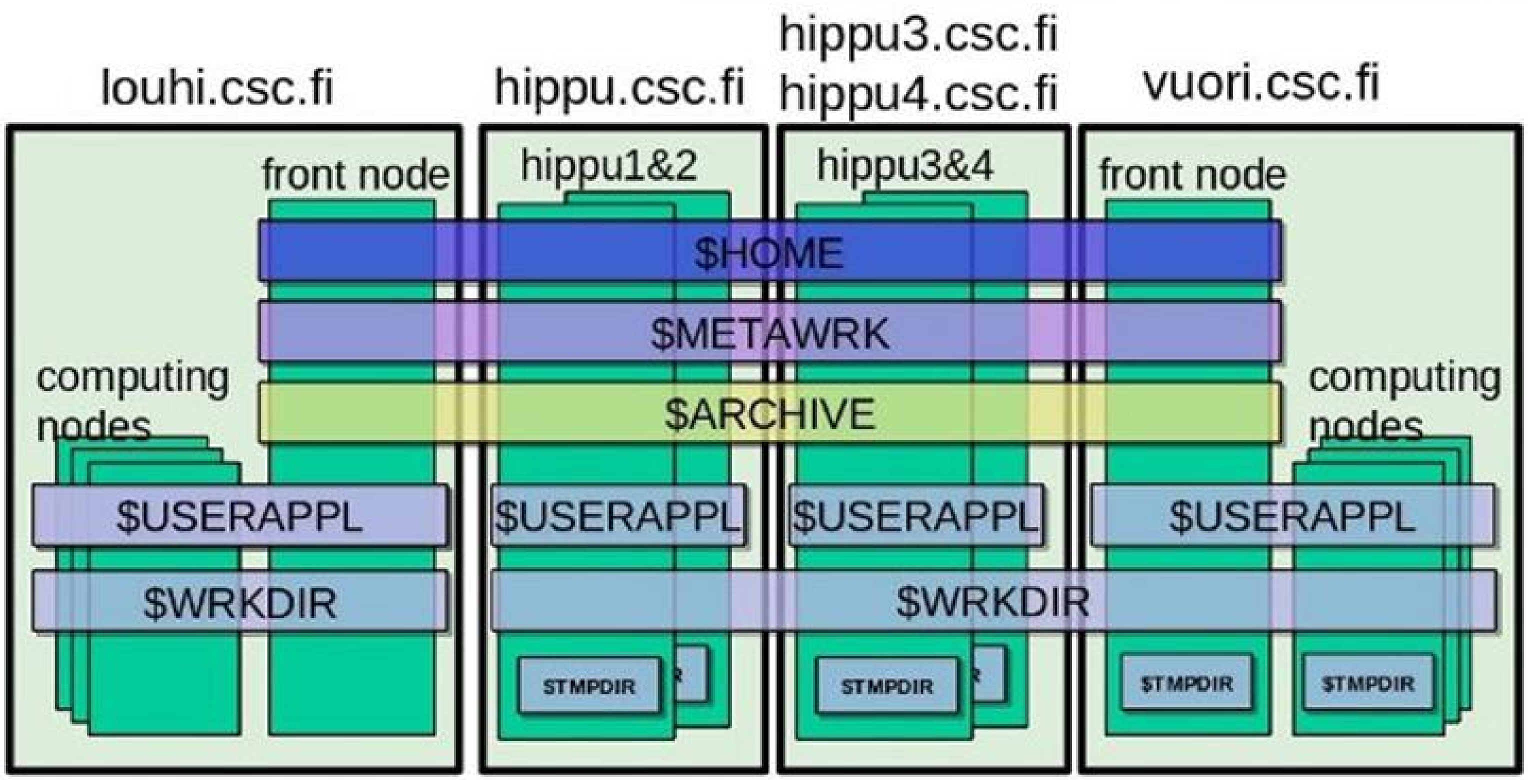
Taito module system

- Old module system has not been actively maintained for a long time
- Robert McClay has written a new and modified version from a scratch
  - More robust
  - Slightly different internal logic
  - Actively developed (bug fixes, new features like support for accelerators, etc.)

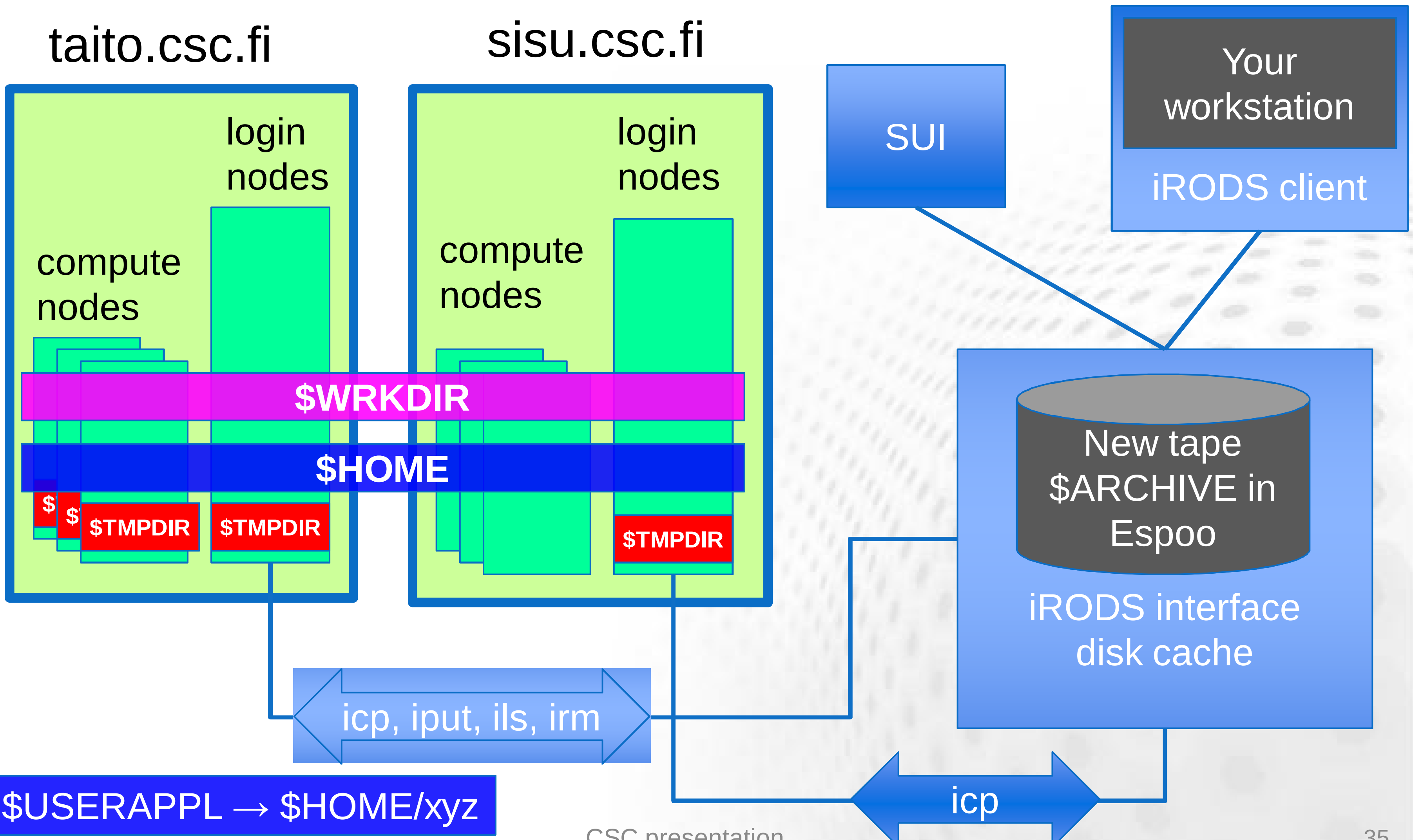
Key differencies (Taito vs. Vuori)

- `module avail` shows only those modules that can be loaded to current setup (no conflicts or extra dependencies)
  - Use `module spider` to list all installed modules and solve the conflics/dependencies
- No PrgEnv- modules
  - Changing the compiler module switches also MPI and other compiler specific modules

Disks at Espoo



Disks at Kajaani



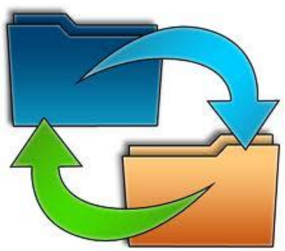
$USERAPPL → $HOME/xyz

Disks

- 2.4 PB on DDN
  - New $HOME directory (on Lustre)
  - $WRKDIR (not backed up), soft quota ~ 5 TB
- $ARCHIVE ~1 - 5 TB / user, common between Cray and HP (@ Espoo)
- Disk space through IDA
  - 1 PB for Univerisities
  - 1 PB for Finnish Academy (SA)
  - 1 PB to be shared between SA and ESFRI
  - Additional 3 PB available later on
- **/tmp** (around 2 GB) to be used for *compiling codes*

Moving files, best practices

- tar & bzip first
- rsync, not scp
  - `rsync -P username@hippu1.csc.fi:/tmp/huge.tar.gz .`
- Blowfish may be faster than AES (CPU bottleneck)
- Funet FileSender (max 50 GB)
  - *https://filesender.funet.fi*
  - Files can be downloaded also with wget



- Coming: iRODS, batch-like process, staging
- IDA
- CSC can help to tune e.g. TCP/IP parameters
  - http://www.csc.fi/english/institutions/funet/networkservices/pert
- FUNET backbone 10 Gbit/s

ARCHIVE, dos and don’ts



- Don’t put small files in $ARCHIVE
  - Small files waste capacity
  - Less than 10 MB is small
  - Keep the number of files small
  - Tar and bzip files
- Time to retrieve (any) file 3 min + 1s/10MB
- Don’t use $ARCHIVE for incremental backup (store, delete/overwrite, store, …)
  - Space on tape is not freed up until months or years!
- Maximum file size 300GB
- Default quota 2 TB per user, new likely up to 5 TB
- New ARCHIVE being installed, consider if you really need all your old files. *Transfer from old to new needed*.

Use profiles

- Taito (HP)
- Serial and parallel upto about 256 cores (TBD)

- Sisu (Cray XE30)
- Parallel up to thousands of cores
- Scaling tests

Queue/server policies

- Longrun queue has drawbacks
  - Shorter jobs can be chained
- Apps that can’t restart/write checkpoint?
  - Code you use to run very long jobs?
- Large memory jobs to Hippu/HP big memory nodes
  - Think about memory consumption
- Minimum job size in Cray
  - Your input?

Documentation and support

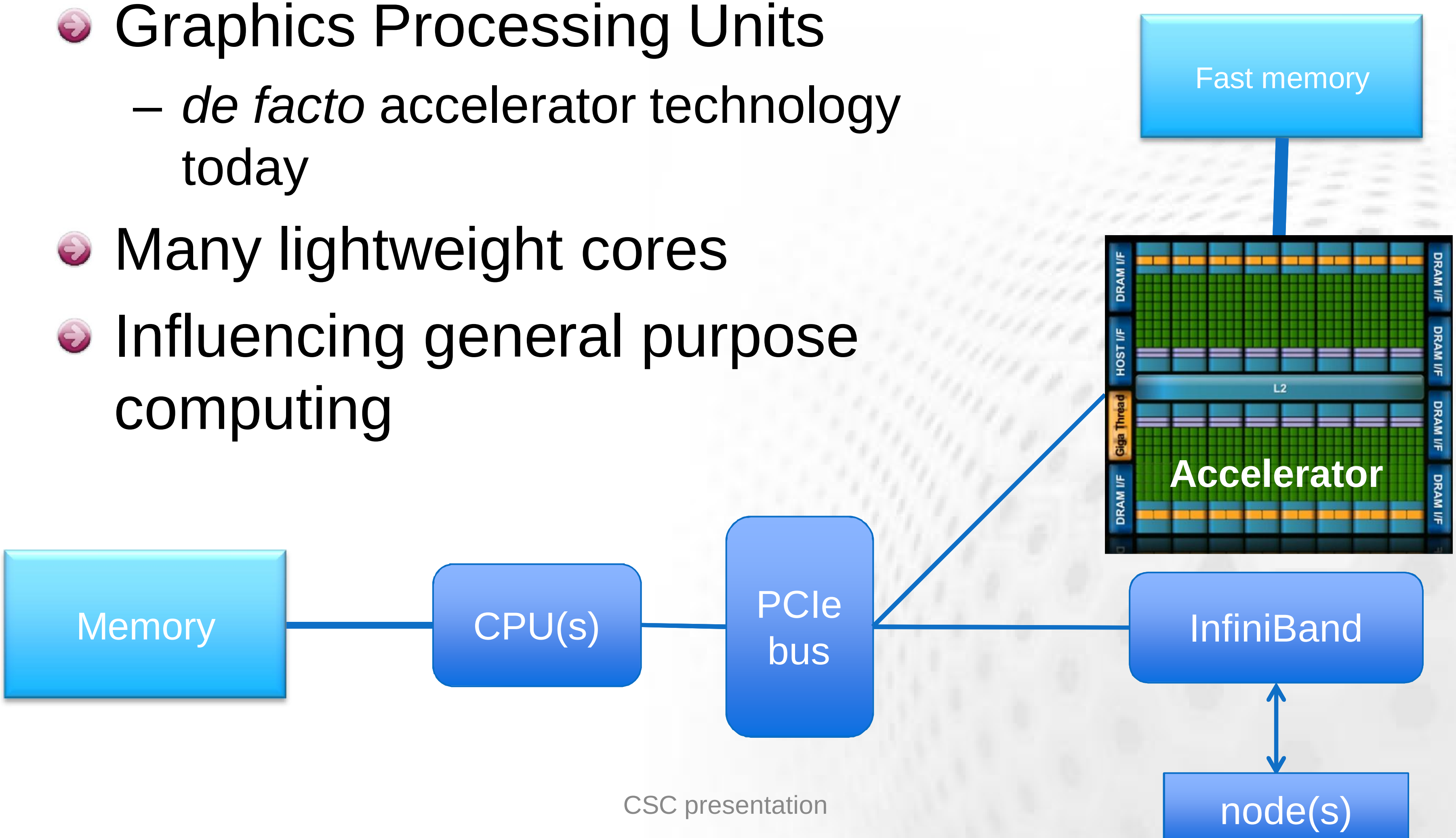
- User manual being built, FAQ here:
  - *https://datakeskus.csc.fi/en/web/guest/faq-knowledge-base*
  - Pilot usage during acceptance tests
- User documetation's link collection
  - *http://www.csc.fi/english/research/sciences/chemistry/intro*
- First HP Workshop materials:
  - http://www.csc.fi/english/csc/courses/archive/taito-workshop
- Porting project
  - All code needs to be recompiled
  - Help available for porting your code
- List of first codes, others added later, some upon request
- User accounts
  - ***HP: recent vuori users moved automatically***
  - ***Cray: recent Louhi users moved automatically***
  - Others: request from **usermgr@csc.fi** with *current* contact information

Grand Challenges

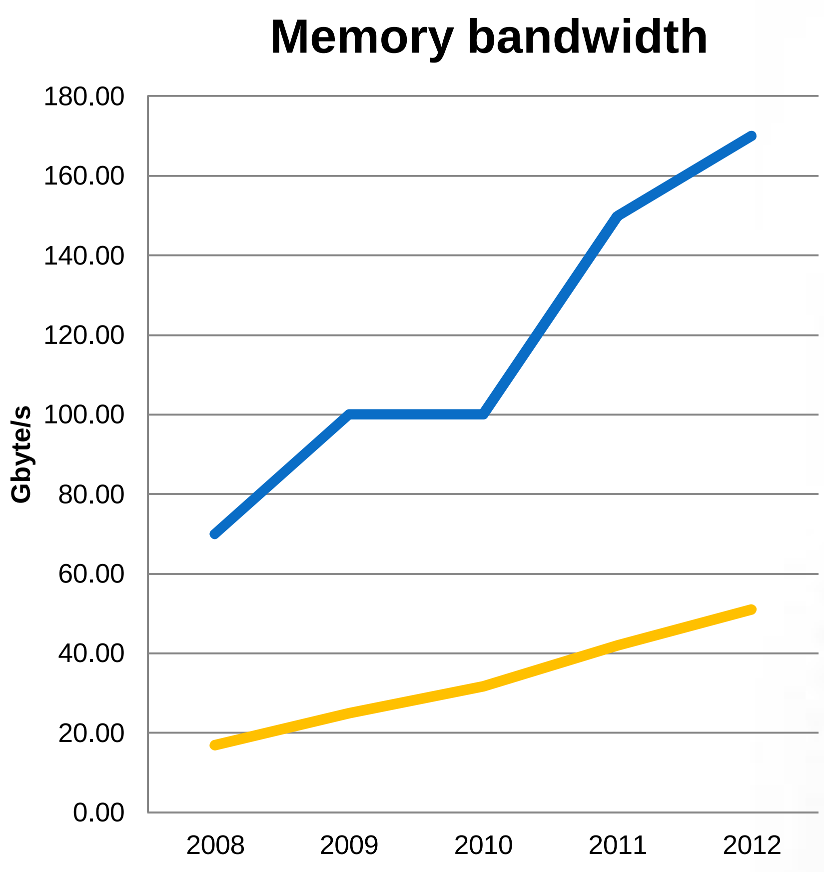
- Normal GC call out in the end of December 2012
  – new CSC resources available for a year
  – no bottom limit for number of cores
- Special GC call (mainly for Cray) out in the end of December 2012
  – possibility for short (day or less) runs with the whole Cray
  – What do you need?
- Remember also PRACE/DECI
  – http://www.csc.fi/english/csc/news/customerinfo/DECI10callopen

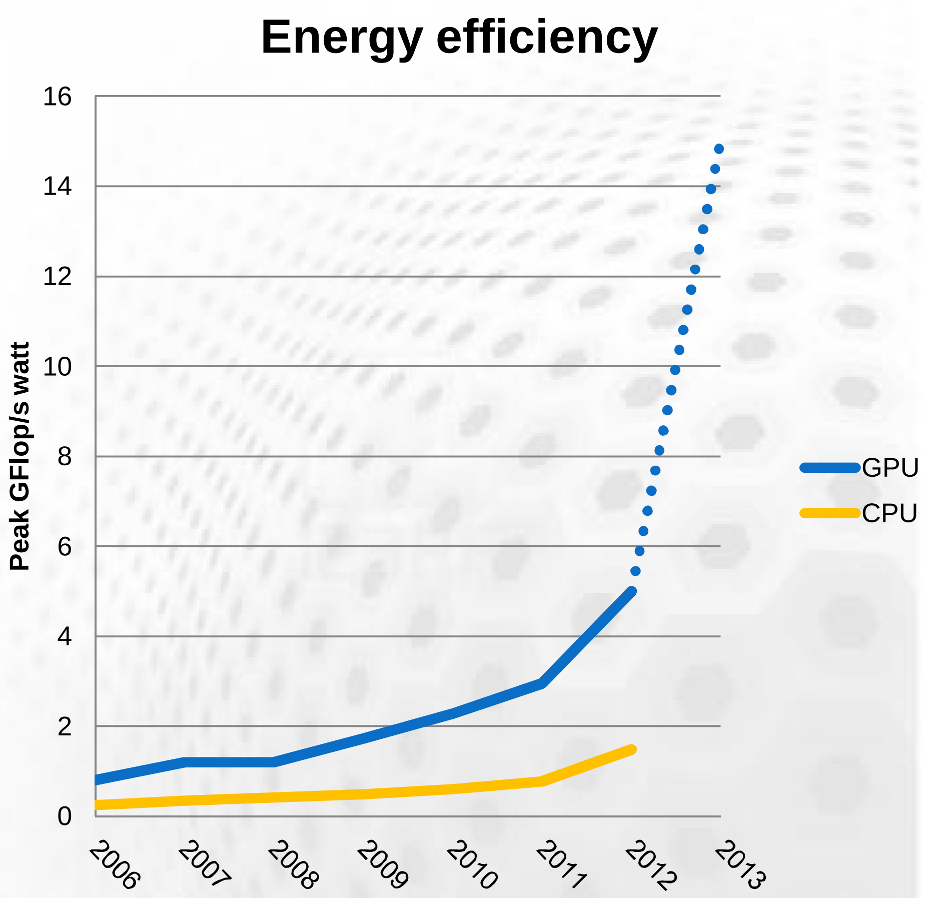
Accelerators

- Add-on processors
- Graphics Processing Units
  - *de facto* accelerator technology today
- Many lightweight cores
- Influencing general purpose computing



Evolution of CPU and GPU performance





Future directions in parallel programming

- **MPI-3** standard being finalized
  - Asynchronous collective communication etc.
- Partitioned Global Address Space (PGAS)
  - Data sharing via global arrays
  - Finally starting to see decent performance
  - Most mature: **Unified Parallel C**, **Co-Array Fortran** (in Fortran 2008), **OpenSHMEM**
- Task Dataflow -based parallel models
  - Splits work into a graph (DAG) of tasks
  - **SmpSs**, **DAGUE**, **StarPU**

CSC RESOURCES AVAILABLE FOR RESEARCHERS

Currently available computing resources

- Massive computational challenges: Louhi
  - > 10 000 cores, >11TB memory
  - Theoretical peak performance > 100 Tflop/s
- HP-cluster Vuori
  - Small and medium-sized tasks
  - Theoretical peak performance >40 Tflop/s
- Application server Hippu
  - Interactive usage, without job scheduler
  - Postprocessing, e.g. vizualization
- FGI



Novel resources at CSC

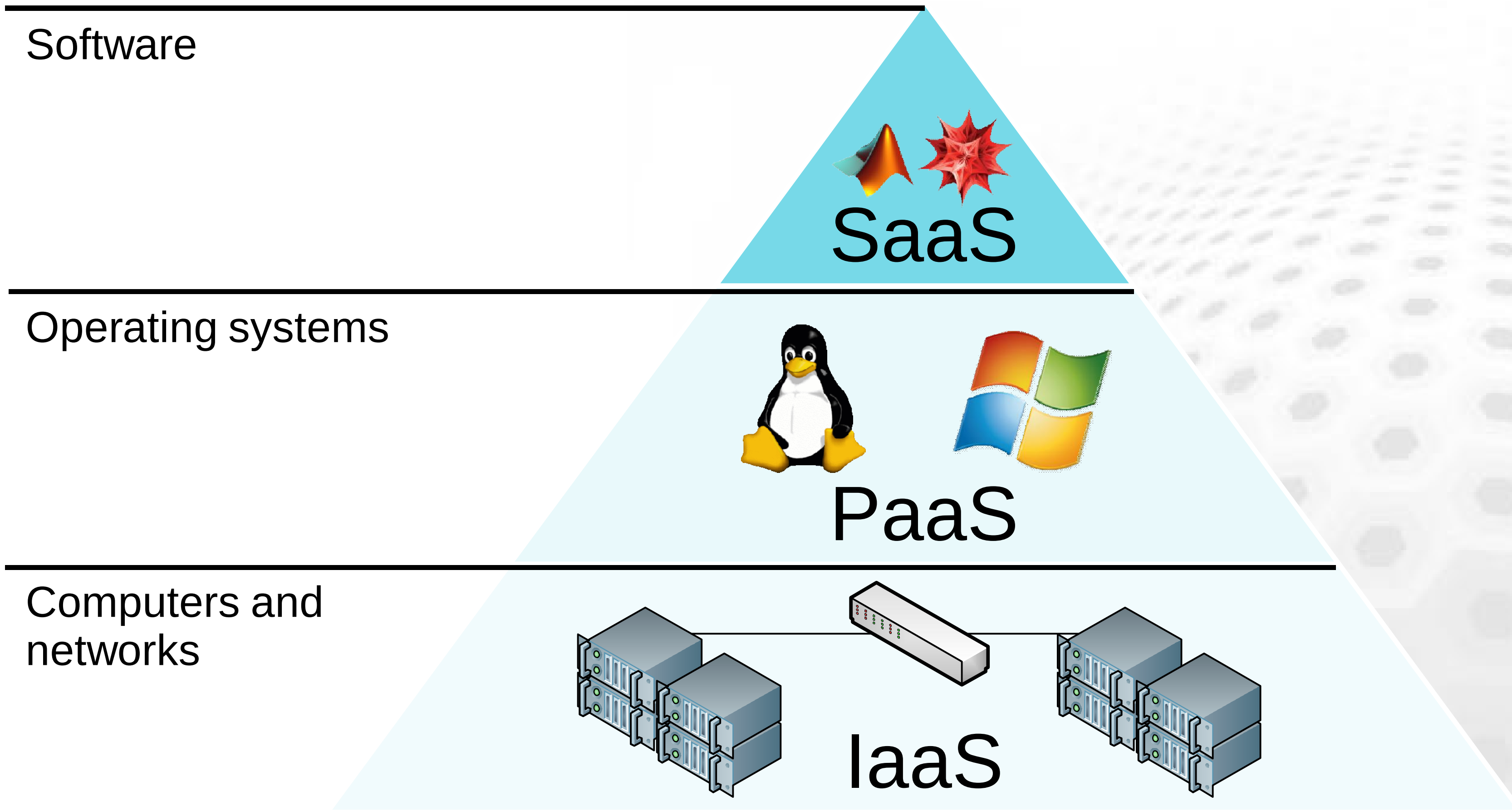
- Production (available for all Finnish researchers)
  - **Vuori**: 8 Tesla GPU nodes
  - **FGI**: 88 GPUs (44 Tesla 2050 + 44 Tesla 2090)
    - GPU nodes located at **HY**, **Aalto**, **ÅA**, **TTY**
- Testing (primarily for CSC experts)
  - **Tunturi:** Sandy Bridge node, cluster
    - Porting to AVX instruction set
  - **Mictest**: Intel MIC prototype node
    - Several beta cards

Old capacity decommissions

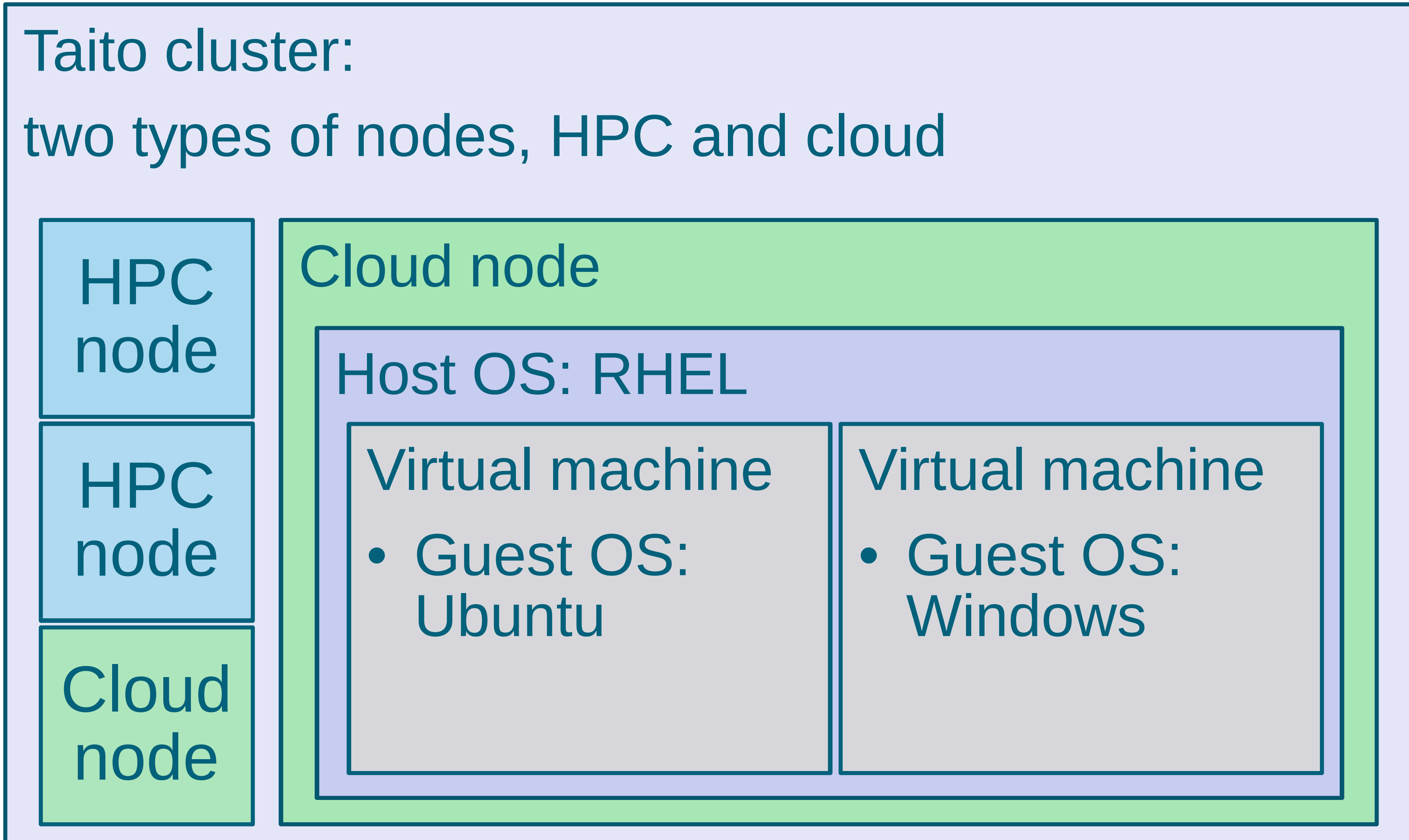
- Louhi decommissioned after new Cray is up and running
  - quite probably fairly short overlap
- Vuori decommission at end of 2013
- Think ahead of data transfers

CSC cloud services

Three service models of cloud computing



Example: Virtualization in Taito



Traditional HPC vs. IaaS

	Traditional HPC environment	Cloud environment Virtual Machine
Operating system	Same for all: CSC's cluster OS	Chosen by the user
Software installation	Done by cluster administrators Customers can only install software to their own directories, no administrative rights	Installed by the user The user has admin rights
User accounts	Managed by CSC's user administrator	Managed by the user
Security e.g. software patches	CSC administrators manage the common software and the OS	User has more responsibility: e.g. patching of running machines
Running jobs	Jobs need to be sent via the cluster's Batch Scheduling System (BSS)	The user is free to use or not use a BSS
Environment changes	Changes to SW (libraries, compilers) happen.	The user can decide on versions.
Snapshot of the environment	Not possible	Can save as a Virtual Machine image
Performance	Performs well for a variety of tasks	Very small virtualization overhead for most tasks, heavily I/O bound and MPI tasks affected more

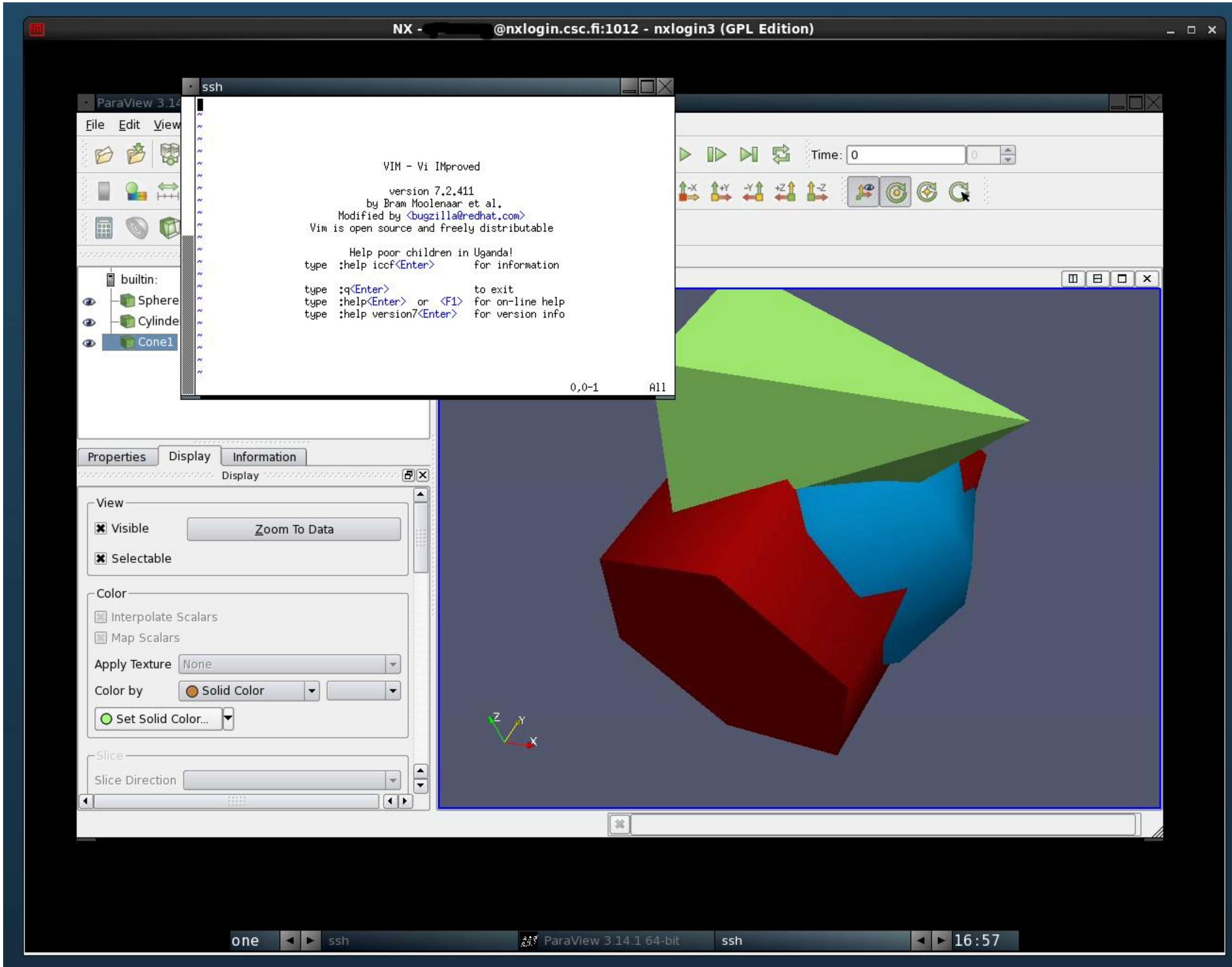
Cloud: Biomedical pilot cases

- Several pilots (~15)
- Users from several institutions, e.g. University of Helsinki, Finnish Institute for Molecular Medicine and Technical University of Munich
- Many different usage models, e.g.:
  - Extending existing cluster
  - Services run on CSC IaaS by university IT department for end users (SaaS for end users)

NX for remote access

- Optimized remote desktop access
  – Near local speed application responsiveness over high latency, low bandwidth links
- Customized launch menus offer direct access CSC supported applications
- Working session can saved and restored at the next login
- Further information: *http://www.csc.fi/english/research/software/freenx*

NX screenshot



Customer training

- Taito (HP)
  - Taito cluster workshop 28.-30.11.2012
    - Materials
  - Next one likely in January 2013
- Sisu (Cray)
  - February 26 - March 1, 2013 (mostly for pilot users, open for everyone)
  - May 14 - May 17 (for all users, a PATC course, i.e. expecting participants from other countries too)

How to prepare for new systems

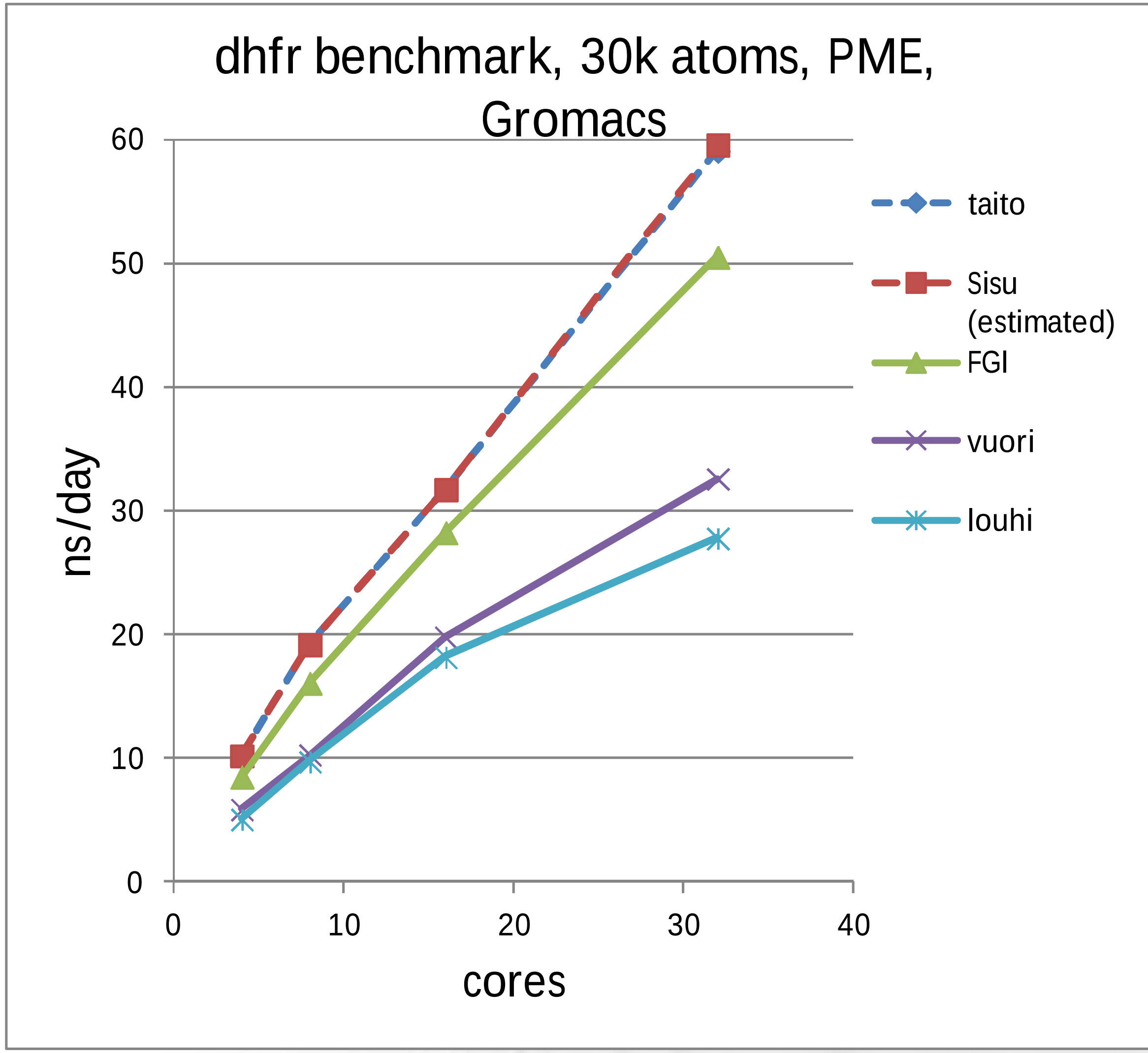
- Participate in system workshops
- Try Intel/GNU compiler in advance, PGI upon request
- Check if your scripts/aliases need fixing (***bash***)
- A lot of resources available in the beginning: prepare ahead what to run!
- The traditional wisdom about good application performance will still hold
  - Experiment with all compilers and pay attention on finding good compiler optimization flags
  - Employ tuned numerical libraries wherever possible
  - Experiment with settings of environment variable that control the MPI library
  - Mind the I/O: minimize output, checkpoint seldom

Sisu&Taito vs. Louhi&Vuori vs. FGI vs. Local Cluster

	Sisu&Taito (Phase 1)	Louhi&Vuori	FGI	Taygeta
Availability	1Q/2Q 2013	Available	Available	Available
CPU	Intel Sandy Bridge, 2 x 8 cores, 2.6 GHz, Xeon E5-2670	AMD Opteron 2.3 GHz Barcelona and 2.7 GHz Shanghai / 2.6 GHz AMD Opteron and Intel Xeon	Intel Xeon, 2 x 6 cores, 2.7 GHZ, X5650	
Interconnect	Aries / FDR IB	SeaStar2 / QDR IB	QDR IB	
Cores	11776 / 9216	10864 / 3648	7308	360
RAM/core	2 / 4 GB 16x 256GB/node	1 / 2 / 8 GB	2 / 4 / 8 GB	4 GB
Tflops	244 / 180	102 / 33	95	4
GPU nodes	in Phase2	- / 8	88	-
Disc space	2.4 PB	110 / 145 TB	1+ PB	0.8 TB

Conclusions

- Performance comparison
  - Per core performance ~2 x compared to Vuori/Louhi
  - Better interconnects enhance scaling
- Larger memory
- Collective communication



Round robin

Nino Runeberg, Atte Sillanpää, CSC – IT Center for Science Ltd.

Round robin

- Your research interest, how CSC can help?
- Queue length: 3 days enough?
  - Codes that can't checkpoint?
  - Other reasons?
- Is memory an issue for you?
  - 256 GB/nodes usage policy?
- Special libraries/tools?
- Special Grand Challenge needs?
- Need to move a lot of files? (from where?)
- Do you need GPGPU/MICs? Which code?